Strategic Exploration and Analysis Using Interactive Visualization of Health Care Databases

Joseph I. Bormel, M.D., M.P.H., G. Octo Barnett, M.D.
Massachusetts General Hospital, Laboratory of Computer Science, Boston, Massachusetts

Effective analysis of health care information is a central component in formulating strategies for improvement initiatives. These initiatives involve the analysis of episodes of care, quality measure comparisons, resource utilization, and investigation of temporal relationships between various factors and outcomes. The ability to efficiently recognize the principal patterns and their variations presents special challenges. This paper describes the design issues and application of the Interactive Visual - Exploratory Data Analysis (IV/EDA) system. IV/EDA is a graphical approach to exploratory analysis of health care information designed to rapidly find and communicate relationships essential to formulating and evaluating strategies that address health care improvement.

INTRODUCTION

Tracking the outcomes of our health care system requires unbiased and rigorous collection of appropriately structured data. Once these observations are available in an electronically readable form, the next task is extracting meaningful information to discover best medical practices from the data. Barriers to this learning include 1) inherent complexities of medical data, 2) the mechanics of database management systems, 3) the appropriate use of statistical concepts and models for exploratory data analysis[1], 4) graphical display techniques[2] that incorporate cognitive psychological issues of large data set presentation[3,4] and 5) the limited professional time and reimbursement for conducting these studies.

Exploratory Data Analysis (EDA) has been described as techniques to make data more easily and effectively handled by those conducting the detective work of extracting the meaningful information[5]. EDA software facilitates unstructured, interactive visual exploration of relationships using multiple graphical displays linked to the data. EDA methods are essential in analyses of large data sets because they can provide perspectives which would otherwise be obscured or completely hidden. EDA methods have been used in conjunction with statistical methods in medical[6] and psychiatric[7] research.

Despite the availability of powerful tools, it is still difficult to evaluate meaningful trends. This difficulty results from inefficient methodologies to explore the data and examine hypotheses generated during this exploration. These deficiencies can be addressed by a variety of techniques.

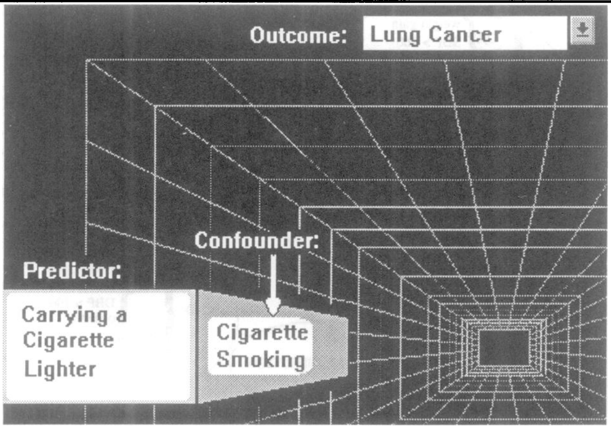

Figure 1: Regression data display. This is a computer generated depiction of the interrelationship of 'Carrying a Cigarette Lighter' and having an outcome of Lung Cancer. The depth queue, labeled 'Cigarette Smoking,' indicates that the apparent predictive power of knowing the patient's cigarette lighter carrying status is diminished by adding knowledge of their smoking status. In fact, once we know smoking status, knowledge of whether a patient carries a cigarette lighter adds no new information about propensity for lung cancer. Visual displays like this have the potential to communicate these kinds of interacting relationships in a database.

Computer programs that provide interactive, visual exploratory techniques for health care relational databases have previously been described[8] The relationships displayed included simple co-occurrence frequencies in a graphical format. This paper describes the continuing development of this effort, called IV/EDA (Interactive Visual Exploratory Data Analysis). It builds on the earlier work by providing more complex frequency of occurrence displays with dynamic grouping, scaling, and focusing. In addition, correlation and regression models, including display of inter-model differences, are supported by visualization techniques. This allows interrelationships to be demonstrated without the user needing to formally hypothesize the relationship.

For example, the single variable *carrying a cigarette lighter* might be highly predictive for subsequently developing lung cancer. However, its predictive power is completely diminished by the knowledge that almost all cigarette lighter carriers also smoke. By displaying the univariate (i.e. considered by itself) significance of *carrying a cigarette lighter* and depicting its loss of significance noted in an exhaustive set of other models including those containing *cigarette smoking*, the loss of predictive ability is made clear when smoking is considered. See **Figure 1**.

METHODS / DESIGN ISSUES

The IV/EDA program is designed to overcome the five previously identified barriers to identifying meaningful information from health data.

Medical Data Issues

Medical databases present special problems not associated with non-medical databases. The significance and interpretation of the medical observations recorded in the database requires special appreciation of uncertainty, confidentiality, domain-specific ambiguity, granularity of observations, missing status, and controlled vocabulary issues. Many special techniques are necessary to cope with these issues including recoding, aggregating, mapping and stratification.

Because of the nature of medical data, a necessary feature is a mechanism to hide the provider and patient identity and any other fields that contain sensitive information not pertinent to the question being addressed. This confidentiality feature is not routinely provided by non-medically oriented software.

The medical context of the data also includes knowledge-based recoding. For example, in one study looking for disease remission, it was helpful to stratify the entry groups according to whether the patients met diagnostic criteria and whether the physician's diagnosis agreed. This stratification could only be achieved by a knowledgeable researcher aware of the possible implications of these variables. Stratification also provides a means to perform risk adjustment in an exploratory setting. Creating an easy to use, rapid environment to form these kinds of groupings is an important design feature of IV/EDA.

Data Access Issues

IV/EDA has specific features and capabilities that enable it to effectively utilize medical database management systems. Using a point-and-click environment with pick lists containing all field names, the user can select data of interest and generate the correct Structured Query Language (SQL) command to calculate and retrieve necessary information. Although this capability is found in most popular database management systems, it is extended in IV/EDA to include special treatment of missing data, aggregation features, and other application-specific query capabilities directed at medical dataset analysis needs. For example, the program generates syntactically correct code for SAS (a statistical software program) to perform data testing such as Chi-Square, T-tests and Survival Analysis.

Statistical Analysis Issues

The appropriate use of statistical concepts and models for exploratory data analysis is a complex issue which has been addressed by other authors[9]. There is no substitute for a fundamental statistical conceptual framework. For example, confusion frequently occurs in recognizing the distinction between the results of exploratory and confirmatory analyses.

To the statistically savvy user, however, IV/EDA provides the ability to more rapidly obtain answers to meaningful questions and concurrently account for confounding and collinearity issues. This feature was demonstrated in the 'Cigarette Lighter/Lung Cancer' example in **Figure 1**.

The ability to use derived statistical properties of database information, such as confounding, is an effective means to reduce the likelihood of interpretation errors. For example, grouping and sorting fields by their correlation with an outcome of interest is useful. This provides another fast way to focus attention from hundreds of variables down to a few, thus eliminating the potential of missing clearly significant variables.

Graphical/Cognitive Design Issues

The judicious use of graphical display techniques con-

Table Data

Dept \ Adverse Event Type	Missed Treatment	Needle Stick	Re-Admit	Medication Error
Med. ICU	39	5	4	33
Med. 5th fl.	0	0	0	48
Surgical ICU	1	0	1	24
Surg. 6th fl.	0	4	14	18
Ob/Gyn	2	5	4	35
Pediatrics	1	2	1	27

Figure 2: Sample adverse event data. The above table show the incidence of four types of adverse events in six departments of a hypothetical hospital.

tributes dramatically to the ability to query the database quickly as well as to present the answers more effectively. It is difficult, if not impossible, to illustrate the value of the dynamic nature of interactive visualization in paper form. Nevertheless, the figures help illustrate the utility of IV/EDA.

Cognitive psychological issues of large data set presentation is a broad topic and the subject of much research. Stephen Kosslyn has written "a picture can be worth a thousand words—but only if you can decipher it." In his book, Elements of Graph Design, he points out how the brain processes visual input quite differently than a camera. The resulting mental impressions of the visual input are in part formed by unconscious filtering. The visual system operates as if it has many distinct input channels, each tuned to different details. An understanding of these filters and channels is essential to convey information effectively.

Another property of the brain is that it can retain only a certain amount of information at any given time. Kosslyn therefore points out that a graph should not require the reader to hold more than four perceptual groups at one time. This creates opportunities for interactive visualization systems to clarify large datasets by focusing on the largest or most important groups.

To illustrate this approach, consider the table of adverse events data in **Figure 2**. This demonstrates numerically the adverse events by type and department for one month at a hypothetical hospital. **Figure 3** shows a visual depiction of this table with the *Surgical ICU* selected as the point of focus. Notice that there are several advantages of this graphic. First, each category of event type and department is depicted by a labeled button. The button is roughly scaled to communicate the total number of events associated with that grouping. By displaying the table of numbers in this way, it becomes immediately clear that *Medication Errors* is the largest and, therefore, most dominant type of adverse event. The department groupings allows the user to notice at a glance that the *Medical ICU* has many more events than the *Surgical ICU.* The lines connecting the *Surgical ICU* with the event types are also scaled such that the thicker line represents more events than the thinner line. These lines are updated to change the focus as the user points and clicks on different labeled buttons.

IV/EDA can provide many other perspectives to the user. For example, although **Figure 3** shows only the *Surgical ICU* data, the data for the other departments can be shown simultaneously using unique colors. This allows easy inspection and direct comparison of each department's adverse events.

One technique used in quality improvement initiatives is to display items sorted from largest to smallest groups in a histogram. This is illustrated in **Figure 4** and is called a Pareto diagram[10]. The points along the ascending cumulative percentage line may be selected within IV/EDA to provide a *Pareto Focus* (as shown in **Figure 5**). For example, point labeled "*18=83%*" means that by focusing in on groups of 18 or more adverse events, 83% of all adverse events are accounted for. By clicking on this label, the co-occurrence depiction is updated, highlighting those departments and event combinations that account for more than 80% of the most frequent adverse occurrences in the hospital. This is shown in **Figure 5.** With large source tables, the value of being able to target one's focus in these ways increases the effectiveness of exploration and clarity of result presentation.

Resource Issues

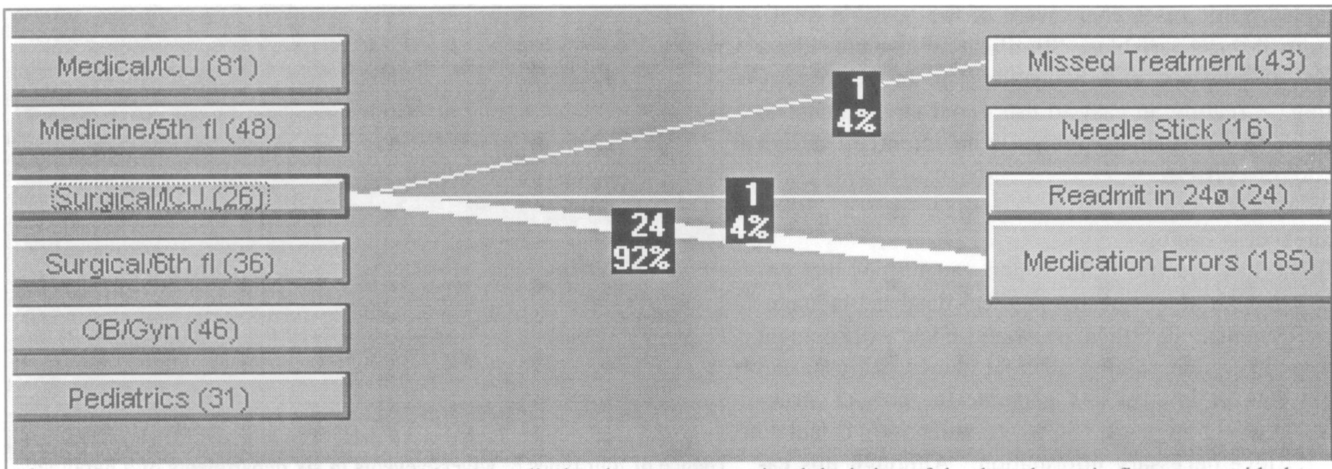

Figure 3: Cross tabulation depiction. This screen display demonstrates a visual depiction of the data shown in figure two with the surgical ICU's perspective selected. Any element of the display can be selected to show other perspectives or combinations of perspectives.

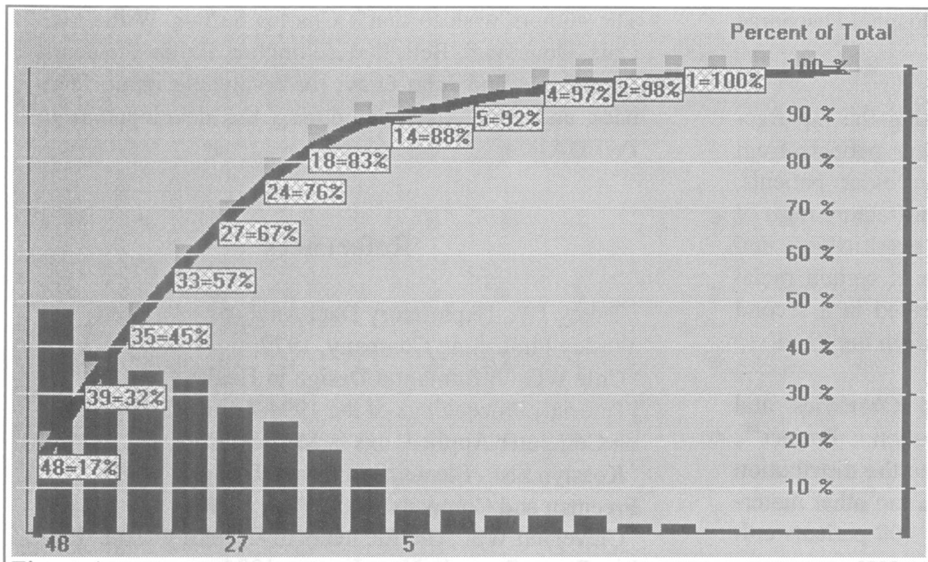

Figure 4: Pareto Diagram. The diagram above shows the same data from figure two. It is grouped with the largest number, the 48 medication errors on the 5th floor shown as the extreme left bar; then successively smaller numbers are shown as the vertical bars extending to rightward. The line extending to the top right shows the cumulative percentage of adverse ends accounted for by including successively smaller groups.

mation. By rapidly focusing on salient issues, IV/EDA can provide the user with analytic tools necessary to effectively address health care problems.

Professional time and research dollars are scarce resources. The increasing availability of information technology and increased focus on cost and quality of care have created demand for better health care data repositories, as well as the enhanced ability to use these repositories.

IV/EDA addresses the latter issue by reducing the time it takes to query and comprehend the resulting infor-

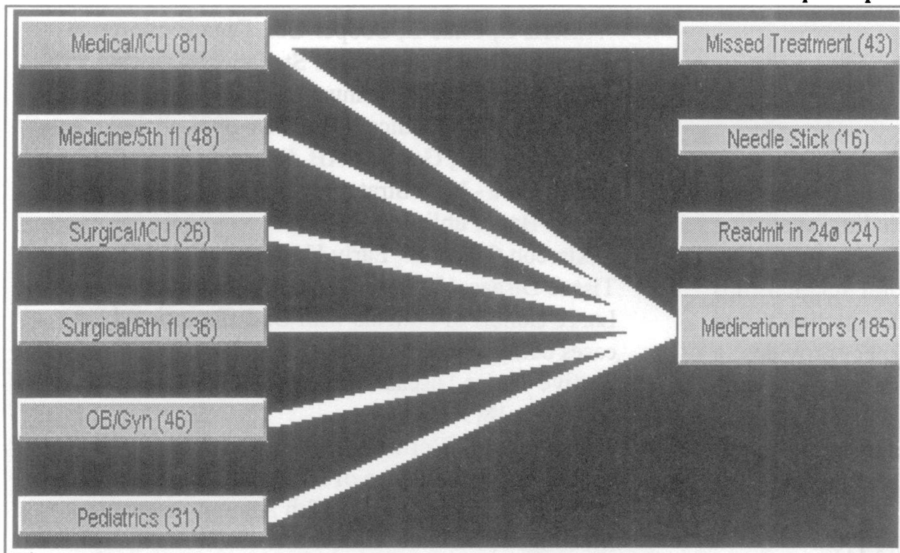

Figure 5: Co-occurrence depiction focused on the 7 largest adverse events groups shown in figure 2. By clicking on the cumulative percent curve at the 80% point on figure 4, the visual display of the table changes to the perspective of the Pareto diagram, highlighting the 'vital few' adverse events worth addressing first. This is referred to as IV/EDA's Pareto Focus.

RESULTS

Evaluative Studies

IV/EDA has been used to explore several large databases for a variety of purposes including:

- Cooperating Systematic Studies of Rheumatic Diseases (CSSRD) Coordinating Center Study of Early Undifferentiated Connective Tissue Disease[11], Salt Lake City, Utah. IV/EDA was used to directly examine patient progression from inception diagnosis to follow-up status (diagnosis, death, or missing). It was also used to look at physician diagnosis and criteria diagnosis concordance and demographic patterns. In addition, IV/EDA was used to perform a retrospective review based on several outcomes to examine the potential prognostic indicators.

  The principal investigator of the study was able to identify a difference in remission rates based on diagnosis stratified by physician agreement with criteria diagnosis within minutes of initial visualization of the large database[8]. Prior to use of IV/EDA, this relationship, as well as others, had not been explored because of the relative inaccessibility of the data associated with conventional tools.

- UCLA Osteoporosis Study. IV/EDA was used to examine racial differences in the incidence of hip fractures over a twenty year period using data from the National Hospital Discharge

Survey as well as the California Hospital Discharge Database[12].

IV/EDA was helpful in aggregating the hip fractures caused by trauma in younger patients from those caused by osteoporosis in older patients, demonstrating racial differences in fracture rates of these older patients using further stratification, and identifying increases in the rates in certain racial groups. These studies were repeated in a second large independent database to confirm the trend.

- Massachusetts General Hospital Obstetrics and Gynecology Clinical Workstation Project[13]. IV/EDA has been used to examine the distribution of pregnancies by parity, gravida, and other maternal information of approximately 400 patients collected during the process of routine care.

IV/EDA has displayed provider characteristics such as anticipated peak demands from the estimated date of confinement variable. Using the *Pareto Focus* feature, specific providers and timeframes were visually presented which depicted periods of anticipated high resource use. This type of information could be used to adjust scheduling when feasible given the external constraints.

Rates of compliance with health maintenance issues, such as recording blood pressure, has also been examined. These have been stratified by care giver and other variables to look for process improvement opportunities.

DISCUSSION

Large dataset analysis is a central component in identifying strategic opportunities to improve health care delivery. This takes many forms including analyzing episodes of care, comparing clinical quality, predicting resource needs, and examining temporal relationships between interventions and outcomes. Interactive visual exploration as described in this paper provides a quicker and more effective way to manage the large amounts of data that underlie these studies. It also provides an enhanced ability to communicate findings to those people involved in the improvement process.

ACKNOWLEDGMENTS

Dr. Bormel is currently a fellow supported by the National Library of Medicine training grant LM7902.

The authors wish to thank Charles Safran, William G. Cole, Douglas S. Bell, Ilan Zellingher, Philip Clements, Fran Cook, and John Orav, for invaluable input, feedback and encouragement during the development of IV/EDA.

References

[1] Tukey JW. Exploratory Data Analysis. Addison Wesley Publishing Company, 1977.
[2] Cole WG. Information Design in Health Care. [Tutorial] Proceedings of the 1994 Annual Symposium on Computer Applications in Medical Care.
[3] Kosslyn SM. Elements of Graph Design. W.H. Freeman and Company, New York. 1994.
[4] Cleveland WS. The Elements of Graphing Data. Hobart Press, Summit, New Jersey. 1994.
[5] Tukey JW. Exploratory Data Analysis. Addison Wesley Publishing Company, Inc, 1977.
[6] Herrmann FR, Safran C. Exploring A Hospital-Wide Database: Integrating Statistical Functions With ClinQuery. Proceedings of the 1991 Annual Symposium on Computer Applications in Medical Care, 583-7.
[7] Politser PE, Berwick KM, Murphy JM, Goldman PA, Weinstein MC. Uncovering Psychiatric Test Information with Graphical Techniques of Exploratory Data Analysis. Psychiatry Research, 1991 Oct, 39(1):65-9.
[8] Bormel JI, Ferguson LR. Visualization and Analysis of Co-occurrence and Cross-tabulation data in Medical Research. Proceedings of the 1994 Annual Symposium on Computer Applications in Medical Care, 944-8.
[9] Aliferis C, Chao E, Cooper GF. Data Explorer: A prototype Expert System for Statistical Analysis. Proceedings of the 1993 Annual Symposium on Computer Applications in Medical Care, 389-393.
[10] Berwick DM, Godfrey AB, Roessner J. Curing Health Care - New Strategies for Quality Improvement. Jossey-Bass Publishers. 1990.
[11] Alarcon GS, et al. Early Undifferentiated Connective Tissue Disease. I. Early Clinical Manifestations in a Large Cohort of Patients with Undifferentiated Diseases compared with cohorts of well established Connective Tissue Diseases. Journal of Rheumatology, 1991 Sep 18(9):1332-9.
[12] Silverman S, Bormel JI, et al. Racial Variation Trends in the Incidence of Hip Fracture. American College of Rheumatology 58th Annual Scientfiic Meeting. 1994.
[13] Chueh HC. A Clinical Workstation for Obstetrics Replaces the Paper Record at MGH. American Medical Informatics Association Spring Congress, 1995.